Citation:
Brown, A. & Maydeu-Olivares, A. (in press). Modeling forced-choice response formats. In Irwing, P., Booth, T. & Hughes, D. (Eds.), *The Wiley Handbook of Psychometric Testing*. London: John Wiley & Sons.

Modeling forced-choice response formats

Anna Brown

Alberto Maydeu-Olivares

Abstract

To counter response distortions associated with the use of rating scales in personality and similar assessments, test items may be presented in so-called 'forced-choice' formats. Respondents may be asked to rank-order a number of items, or distribute a fixed number of points between several items – therefore they are forced to make a choice. Until recently, basic classical scoring methods were applied to such formats, leading to scores relative to the person's mean (ipsative scores). While interpretable in intra-individual assessments, ipsative scores are problematic when used for inter-individual comparisons. Recent advances in estimation methods enabled rapid development of item response models for comparative data, including the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011a), the Multi-Unidimensional Pairwise Preference model (Stark, Chernyshenko & Drasgow, 2005), and others. Appropriate item response modeling enables estimation of person scores that are directly interpretable for inter-individual comparisons, without the distortions and artifacts produced by ipsative scoring.

Keywords: forced-choice format, ipsative data, single-stimulus format, multidimensional IRT, comparative judgment, dominance models, ideal-point models, unfolding

What are forced-choice response formats?

When thinking of a 'typical' questionnaire, we imagine a series of stimuli (statements, questions, words or phrases, pictures, etc.) that our subjects respond to, usually by selecting one of several response options. For example, we may ask respondents to indicate the extent to which the following statements are true of them:

	very untrue of me	somewhat untrue of me	somewhat true of me	very true of me
I am relaxed most of the time		X		
I start conversations				X

The common characteristic of this type of questionnaire format is that the stimuli (we will refer to them as questionnaire *items*) are responded to **one at a time**. Regardless of the exact type and number of response options used, respondents are supposed to consider one single stimulus at a time, and provide a response to it independently of other stimuli. This response format is called the *single-stimulus format*. The single-stimulus format is by far the most popular in psychometric practice. In his Theory of Data (1964), Clyde Coombs devoted a whole class (Type II) to single-stimulus data.

There is, however, an alternative way to gather responses to the same stimuli. Instead of presenting them one by one, we can present two stimuli together and ask the respondents to express their **preference** with respect to the stimuli presented. For instance, for the two statements from the example above, we can ask them to indicate which statement is **most** true:

	most true of me
I am relaxed most of the time	
I start conversations	X

In this task, the respondent is forced to make a choice (hence the name – *forced-choice format*). Regardless of whether both statements are true or untrue of the respondent, he/she will have to select one that is most true. This will unambiguously imply that the remaining statement is less true; therefore, a full rank order of the two statements is obtained. Examples of tests using forced-choice pairs are the Tailored Adaptive Personality Assessment System (TAPAS; Drasgow, Chernyshenko, and Stark 2010), and the Navy Computerized Adaptive Personality Scales (NCAPS; Schneider et al. 2007).

We can add another statement to the forced-choice pair, and ask respondents to rank order three statements according to the extent to which they are true of them, for example:

	rank order
I am relaxed most of the time	3
I start conversations	1
I catch on to things quickly	2

Another variation of rank ordering is to ask respondents to select only the top and bottom ranks, as follows:

	most / least true of me
I am relaxed most of the time	least
I start conversations	most
I catch on to things quickly	

In the example scenario with three statements, making the ‘most’-‘least’ choices will unambiguously place the remaining statement in-between the two selected statements, and therefore this format is equivalent to complete rank ordering. If, however, we add another statement to make a block of four forced-choice items, selecting the top and bottom ranks would yield an incomplete ranking:

	most / least true of me	rank order
I am relaxed most of the time	least	4
I start conversations		
I catch on to things quickly		
I sympathize with others' feelings	most	1

In the above example, the ranking is *incomplete* because we do not know which of the two remaining statements would receive rank '2', and which one would receive rank '3'. There are more examples of forced-choice formats producing incomplete rankings, for instance this would occur if the respondent were asked to select only the top ranking item from a block of three or more items (as we already know, this format would yield a complete ranking if only two items were involved).

All the above examples conform to an *ordinal* forced choice format (Chan 2003), since only the order of the items is obtained. Examples of tests using ordinal forced-choice formats are the Occupational Personality Questionnaire (OPQ32i; SHL 2006), the Personality and Preference Inventory (PAPI; Cubiks 2010), the Customer Contact Styles Questionnaire (CCSQ 7.2; SHL 1997), the Gordon's Personal Profile Inventory (GPP-I; Gordon 1993), the Survey of Interpersonal Values (SIV; Gordon 1976), and the Kolb Learning Style Inventory (Kolb & Kolb 2005).

Another, more complicated type of ranking is the so-called Q-sort (Block 1961), whereby respondents perform ranking with ties. In this format, respondents have to assign a number of items to several categories, complying with a pre-defined distribution, (i.e. the number of items to be assigned to each category is fixed). For example, respondents are asked to sort 45 items into five piles, according to the extent to which the items are characteristic of them, and to make sure that every pile contains the precise number of items specified:

	Pile Number					
most uncharacteristic	1	2	3	4	5	most characteristic
	5	10	15	10	5	
	Number of items in pile					

Although seemingly a rating task, the Q-sort is in fact a pure forced-choice format, because rating decisions have to consider assignments to categories both below and above, thus necessitating direct comparisons between items. Examples of tests using this format are the California Adult Q-Set (Block 2008), the Riverside Behavioral Q-sort (Funder, Furr and Colvin 2000) and the Riverside Situational Q-sort (Wagerman and Funder, 2009).

Some forced-choice formats go beyond simple rank ordering and involve judgments of **the extent** of preference for one item over another. In our example with two items, respondents may be asked to indicate to what extent one statement is more (or less) true of them:

	much more true of me	slightly more true of me	slightly more true of me	much more true of me	
I am relaxed most of the time				X	I start conversations

Now we not only have the preferential ordering of the items; we also have some quantitative information about the relative merits of the two items. We can also collect quantitative information by asking the respondents to distribute a fixed number of points between several items. For instance, we may ask respondents to distribute 10 points between three statements according to the extent to which the statements are true of them:

	Points (10 in total)
I am relaxed most of the time	1
I start conversations	5
I catch on to things quickly	4

Comparing the above format with the most-least selections using the same items, it is clear that although the preference for the statement 'I start conversations' remains, the latter format captures more information about the extent of that preference. From the example responses it can be seen that the respondent judged 'I start conversations' is only slightly more true of the self than 'I catch on to things quickly'; however, he/she judged it to be much more true of the self than 'I am relaxed most of the time'. Another typical example of this format is asking respondents to distribute 100 points between several items; in this case, the points awarded to each item can be interpreted as percentages of a total amount. This type of forced choice is referred to as *multiplicative* or compositional (Chan 2003).

Having considered several examples of forced-choice formats, the reader will see that they are fundamentally different from single-stimulus formats. When using single-stimulus formats, the respondents make *absolute judgments* about every individual item. When using forced-choice formats, the respondents engage in *comparative judgments*. The Theory of Data (Coombs, 1964) devotes two whole classes (Type I – Preferential Choice, and Type III – Stimulus Comparison) to data obtained by using forced-choice formats[i].

The advantages of presenting questionnaire items using the forced-choice format

While it is clear that forced-choice formats are different from single-stimulus formats, we have not yet discussed why we might want to present questionnaire items in this fashion. Do comparative judgments have any advantages over absolute judgments? Indeed, the forced-choice format does have its own merits.

Firstly, comparative judgments eliminate any systematic response sets that apply uniformly across items (Cheung and Chan 2002). For instance, having to make a choice between items will make it impossible to endorse them all indiscriminately (so-called *acquiescence* bias). It will also make it impossible to produce uniformly elevated or decreased judgments across all items, therefore eliminating rater effects such as *leniency* / *severity*. Indeed, if instead of rating several employees on the same performance indicator, a rater simply rank orders the employees according to their performance, any effects associated with leniency (or severity) of this particular rater are removed. Another example of uniform response sets is the individual tendency to provide extreme versus middle-ground ratings (*extremity* / *central tendency* responding). These tendencies are also overcome by the use of the forced-choice formats.

Secondly, simple rank ordering removes the need for any rating scale. This is an important and perhaps overlooked advantage of **ordinal** forced-choice formats. In addition to the problem of extremity/central tendency, and biases caused by the order of response options and whether they have numerical or verbal anchors (e.g. Schwarz et al. 1991), there are further complex and difficult to classify issues which concern the idiosyncratic interpretation and use of rating scales. Respondents may interpret the response options or any other verbal and non-verbal anchors provided with the rating scale differently, resulting in a violation of the main assumption of rating scales – that their effects are fixed across respondents (Friedman and Amoo 1998). Some argue that performing direct comparisons between items may be simpler for respondents than arriving at ratings for each item – a process that requires discrimination between various response options (Maydeu-Olivares and Böckenholt 2008). This might be particularly true when rating scales contain many response options and provide few verbal anchors, or when the response options are ambiguous or inappropriate for the items. This argument, however, can be turned on its head when we consider extremely

complex forced-choice tasks such as Q-sorts. Clearly, the cognitive complexity of such forced-choice designs is by far greater than that of almost any rating scale.

Thirdly, forced-choice formats may be useful in research involving personal or sensitive information. Chan (2003) provides an example whereby respondents are asked to indicate the proportion of their household income spent on several commodity groups (food, clothes, electricity, etc.). In such a survey, the respondents may not want to disclose their spending patterns in monetary value, but they may be happy to provide a percentage breakdown. Taking issues of privacy into account is important in many research topics; and ensuring that the respondents' privacy is protected may be essential for guaranteeing valid responses.

Fourthly, imposing forced choices directly tackles the problem with lack of differentiation in ratings. This problem, reported over 100 years ago by Thorndike (1920) and still widespread in ratings of individuals, organizations and services, is known as the *halo* effect. The halo effect is characterized by overgeneralized assessments of all characteristics of the rated object based on one important dimension. Unlike uniform rater biases such as leniency/severity, the halo effect might not act uniformly across all items, and instead might influence some of them more than others. Forced choice makes raters differentiate between various characteristics of the rated object, and thus this reduces halo effects. Bartram (2007) showed that the use of forced-choice formats in line manager assessments of employees' job competencies could increase correlations with external measures by as much as 50% compared to single-stimulus competency ratings.

Finally, and perhaps most controversially, forced-choice formats have been used in personality and similar assessments when responses were thought to be subject to *socially desirable responding*, whether due to self-deception or motivated *impression management* (Zerbe and Paulhus 1987). The latter types of distortion often referred to as *faking good* are

of particular concern in high stakes assessments, such as assessments in recruitment or selection of employees. It has been argued that because it is impossible to endorse all items, when combining equally desirable items in the same forced-choice block that this would prevent respondents from endorsing the desirable items and rejecting the undesirable ones. Over the years, some authors have reported positive evidence for the use of forced-choice formats in these contexts (Christiansen, Burns, and Montgomery 2005; Jackson, Wroblewski, and Ashton 2000; Martin, Bowen, and Hunt 2001; Vasilopoulos et al. 2006), while others have found the forced-choice format to be as susceptible to faking as the single-stimulus format (Heggestad et al. 2006). While inarguably dependent on particular questionnaire designs and specific contexts, these disparate findings highlight problems in using the forced-choice format for eliminating motivated response distortions. Matching items on social desirability levels is one such problem. Research shows that respondents have different perceptions of item desirability (Kuncel, Goldberg, and Kiger 2011). Coupled with the fact that item desirability depends on the testing context (e.g. the job for which respondents are being assessed), it follows that even the most careful matching cannot make all respondents in all contexts perceive all items in one block as equally desirable. Another problem was summarized by Feldman and Corah (1960), who argued that direct comparisons between items might actually invite finer distinctions between their desirability levels than rating each item independently, therefore potentially heightening the social desirability effects. The jury is still out on the question of the effectiveness of forced-choice formats in reducing socially desirable responding. We believe that to answer this question conclusively, a better understanding of the psychological process behind socially desirable responding is required, and of how this process may interact with comparative judgments.

Having discussed the potential advantages of forced-choice formats, the reader might wonder why their use is not more widespread. It turns out that the forced-choice formats also

have major disadvantages, which are concerned with scaling. We turn to the scaling of forced-choice data next.

Scaling of forced-choice responses

The purpose of gathering responses on questionnaire items in psychometric applications is to scale the objects of assessment on the psychological attributes measured by the questionnaire. Typically, we are interested in absolute scaling so that every object is associated with a number on a psychological continuum, and objects can be assessed relative to the scale origin. How do we derive absolute scale scores (e.g. personality trait scores,) from relative information provided by forced-choice formats (e.g. relative preferences for personality test items)? This section describes popular scaling methods for forced-choice responses.

A classical approach, or measurement by 'fiat'

Traditionally, forced-choice data has been treated in a similar way to rating scales. Points are awarded for preferring an item; and the points are added to the scale that the item is designed to measure. For instance, preferring an item from a pair may result in one point being added to that scale which is measured by the item:

	most true of me	score
I am relaxed most of the time		0 (to Emotional Stability)
I start conversations	X	1 (to Extraversion)

In blocks of three, four or more items, inverted rank orders (or some linear function of these inverted rank orders) would be added to the respective scales that the items are designed to measure, so that the item ranked first would earn most points and the item ranked last would earn least points:

	most / least true of me	rank order	score
I am relaxed most of the time	least	3	0
I start conversations	most	1	2
I catch on to things quickly		2	1

If an incomplete ranking were obtained, the same logic would apply, with the item ranked first earning most points, the item ranked last earning least points, and the items not ranked earning an equal number of points in between:

	most / least true of me	rank order	score
I am relaxed most of the time	least	4	0
I start conversations			1
I catch on to things quickly			1
I sympathize with others' feelings	most	1	2

For tasks in which participants are asked to distribute a fixed number of points between several items, these points would typically be awarded to the respective scales.

One common feature of all these scoring protocols is that the number of points awarded is quite arbitrary – it is not justified theoretically or empirically. Torgeson (1958) called this type of scaling *measurement by fiat*. Another common feature is that in all the above examples the total number of points awarded in the block is **constant** for all respondents. To illustrate, consider our previous example of a forced-choice pair, and another forced-choice pair measuring the same personality traits, with the following responses from respondent A:

Respondent A	most true of me	score
I am relaxed most of the time		0 (to Emot. Stability)
I start conversations	X	1 (to Extraversion)

Respondent A	most true of me	score
I rarely get irritated		0 (to Emot. Stability)
I make friends easily	X	1 (to Extraversion)

The responses above will result in 2 points being allotted to Extraversion and 0 points to Emotional Stability – 2 points in total. Now consider the following responses to the same item-pairs (respondent B):

Respondent B	most true of me	score
I am relaxed most of the time	X	1 (to Emot. Stability)
I start conversations		0 (to Extraversion)

Respondent B	most true of me	score
I rarely get irritated	X	1 (to Emot. Stability)
I make friends easily		0 (to Extraversion)

These responses will result in 0 points being allotted to Extraversion and 2 points to Emotional Stability – again totaling 2 points.

An immediate implication is that, regardless of the choices made, every participant will receive the same total number of points on the test but that the points will be distributed differently between different scales. This type of data is called *ipsative* data. The name was coined by Cattell in 1944, from the Latin ‘ipse’ (he, himself), reflecting the fact that any score obtained on an ipsative test is relative to self. Indeed, while respondents A and B obtained the same number of points, the points were allotted differently to the two scales. All we can conclude from these scores is that respondent A’s score on Extraversion is higher than his score on Emotional Stability, and that the opposite is true for respondent B. We, however, cannot tell where respondents A and B stand in relation to each other on either of these scales – it is entirely possible, for example, that respondent B is just as high on Extraversion as

respondent A, and he is higher on Emotional Stability still. There is simply no information in the assigned points in order to make such inferences.

To summarize, in ipsative data the number of points on any scale for an individual can only be interpreted in relation to the number of points on other scales, because the total number of points on all scales is the same for everyone. The self-referenced nature of ipsative scores means that scores are not interpersonally comparable. This is because a score on scale A is not scaled in relation to some absolute scale origin, but instead in relation to the individual's own mean. Unlike a *normative* measure, which allows individuals to have different test means, in an ipsative measure the mean is the same for everyone.

Ipsative data have been criticized for the lack of interpretability of individual differences, and other psychometric challenges that follow from this basic property. Clemans (1966) provided a full mathematical account of problems with using ipsative data (see also Dunlap and Cornwell 1994). Other authors have illustrated how these challenges may manifest themselves in applications (e.g. Hicks 1970; Johnson, Wood, and Blinkhorn 1988; Tenopyr 1988; Closs 1996; Meade 2004). In the following section, we discuss psychometric properties of ipsative data.

Psychometric properties of ipsative data

1. Ipsative scores are interpersonally incomparable

This is the most fundamental property of ipsative scores following directly from their definition. Because the same total number of points is allocated to everyone, it is impossible to achieve a high score on one scale without reducing scores on other scales. A high score received might not be a reflection of a high standing on a theoretical attribute in any normative sense (i.e. with reference to a population); instead, it might be simply an artifact of having low scores elsewhere on the test. Fundamentally, people with an identical relative

ordering of attributes will obtain identical ipsative scores, regardless of their normative standing in relation to each other. Clearly, this can have serious implications for assessment decisions in applied settings.

2. Ipsative scores distort construct validity

Because the total test score on ipsative measures is the same for everyone, it has zero variance. Therefore, all elements of the scales' covariance matrix sum to zero (Clemans 1966), and the average off-diagonal covariance is a negative value. In an ipsative test consisting of *k* scales with equal variances, the average correlation among the scales must be

$$\bar{\rho} = -1/(k-1). \quad (1)$$

That is, regardless of the expected relationships among the attributes measured by the test, their measured scales will correlate negatively on average. Clearly, when the true scores on the attributes correlate positively, the ipsative scores will distort these relationships.

Because elements of the ipsative variance-covariance matrix sum to zero, one of the eigenvalues must be zero and maximum likelihood factor analysis cannot be applied (Dunlap and Cornwell 1994). Principal components analysis can be performed on ipsative scores; however, the resulting components are difficult to interpret as they typically consist of scales representing contrasting choices (Cornwell and Dunlap 1994; Baron 1996). Overall, ipsative data compromises the construct validity of forced-choice tests.

3. Ipsative scores distort criterion-related validity

As the variance of the total test score is zero, covariances of the ipsative scales with any external measure will sum to zero (Clemans, 1966; Hicks, 1970). Therefore, any positive covariances with the external variable have to be compensated by some spurious negative covariances, and vice versa (Johnson, Wood, and Blinkhorn 1988). Such distortions to the covariance patterns will be larger when the scales are expected to co-vary with the criterion

mostly positively (or negatively). Overall, ipsative data compromises the criterion-related validity of forced-choice tests.

Partially ipsative data

Having provided a brief summary of the psychometric properties of ipsative data, we show how varying item polarity and point assignment can yield data in which the individual total score on the test is only partially constrained. Consider our previous example of two forced-choice pairs. This time, the second item-pair is designed to indicate Extraversion and the negative end of Emotional Stability (i.e. Neuroticism). Here are possible responses from respondent A:

Respondent A	most true of me	score
I am relaxed most of the time		0 (to Emot. Stability)
I start conversations	X	1 (to Extraversion)

Respondent A	most true of me	score
I worry about things*	X	–1 (to Emot. Stability)
I feel at ease with people		0 (to Extraversion)

The item 'I worry about things' indicates Neuroticism (the opposite end of emotional stability) and is *negatively keyed* (it is scored –1 if preferred in the comparison). The responses in this example will result in the allocation of 1 point to Extraversion and 1 point to Emotional Stability, 0 points in total.

Now consider the following responses to the same item-pairs (respondent B):

Respondent B	most true of me	score
I am relaxed most of the time	X	1 (to Emot. Stability)
I start conversations		0 (to Extraversion)

Respondent B	most true of me	score
I worry about things*		0 (to Emot. Stability)
I feel at ease with people	X	1 (to Extraversion)

These responses will result in the allocation of 1 point to Extraversion and 1 point to Emotional stability, 2 points in total. The reader can see that respondents A and B obtain different numbers of points now. This is because the test contains both positively and negatively keyed items indicating the same attribute. Such tests will yield data that shows some variation in the total score. This variation is still constrained, because in both item-pairs the items indicating Emotional Stability are scored relative to a baseline score obtained on the Extraversion items. The type of data derived from such tests is called *partially ipsative*. It is generally less problematic than fully ipsative data; however, its psychometric problems are not eliminated but merely reduced.

An Item Response Theory approach, or measurement by modelling

The general problem with measurement by fiat is that the scores assigned to items are not justified theoretically or empirically. Specifically, the classical approach to scoring of forced-choice items treats rankings (relative information) as if they were ratings (absolute information). Furthermore, scale scores are computed as the sum of item scores in both the forced-choice and the single-stimulus formats. These two formats, nevertheless, reflect very different response processes and assign very different meanings to item responses.

In the single-stimulus format, a respondent selects a rating option that represents the level of an attribute or behavior which is closest to that of his or her own. The response, therefore, can be directly referenced in relation to a criterion (e.g. the behavior is not demonstrated at all, demonstrated to a small extent, or to a large extent), and consequently to

a scale origin. We make the same implicit assumption when assigning scores to forced-choice items using the classical approach, completely inconsistent with the fact that forced-choice responses represent items' relative positions (in that particular forced-choice block), not the absolute levels on the attributes. Consider the situation when a respondent gives the top rank to a particular item in a block. This choice does not reflect his/her degree of agreement with the item; it simply reflects that he/she agrees with it *more* than with the other items in the block. Thus, similar items indicating the same attribute might be ranked highest in one block and lowest in another, depending on the items they are contrasted against, resulting in item scores that are inconsistent with the implicit assumption of absolute scaling. Taking this reasoning further, we can see that the idea that adding relative positions together somehow will constitute a scale score that reflects an absolute attribute score is also illogical.

Another implicit assumption underlying the classical approach to scoring is that item responses are independent from each other, after they have been controlled for the attribute levels (local independence assumption). This assumption is clearly violated, because items within a block are **not assessed independently** but in relation to each other. Making forced choices creates mutual dependencies between responses to all items in the block.

To summarize, the classical method of assigning scores to forced-choice items does not correspond to the **meaning** of item responses. In other words, the implicit model underlying this scoring bears no relation to the psychological process used in comparative judgments (Meade 2004). The consequences of this misrepresentation in the implicit scoring model are the problems of ipsative data (Brown and Maydeu-Olivares 2013).

Can the situation be improved by adopting an approach that considers the meaning of preferential choices? It turns out that it can. By considering the response process to forced-choice items and devising a model to describe this process, model-based measurement may be inferred for forced-choice data. Mellenbergh (2011, p. 188) named this type of scaling

‘*measurement by modeling’*, in contrast to measurement by fiat. This approach is also known under the name of Item Response Theory (IRT).

Several item response models have been developed recently to infer measurement from forced-choice data. Four such approaches are briefly described in this chapter: the Thurstonian IRT model (Brown and Maydeu-Olivares 2011a), the Zinnes-Griggs model for Unidimensional Pairwise Preferences (Zinnes and Griggs 1974), the Multi-Unidimensional Pairwise-Preference (MUPP) model (Stark, Chernyshenko, and Drasgow 2005), and the McCloy-Heggestad-Reeve (2005) unfolding model for multidimensional ranking blocks. Each of these models has distinct objectives and assumes different forced-choice designs and different properties of items used. All models may be used for creating and scoring **new** forced-choice assessments. The Zinnes-Griggs model can also be applied to estimate item parameters in forced-choice item-pairs where items measuring the same scale are compared (unidimensional items). The Thurstonian IRT model is currently the only one that can be readily applied to data collected with **existing** multidimensional forced-choice questionnaires, with the additional objectives of estimating item parameters and relationships between the latent attributes.

In what follows, we briefly describe these models using a common framework. Specifically, we distinguish two components of models for forced-choice data: 1) a model for the decision process leading to selection of items, and 2) a model for relationships between items and the underlying attributes they measure. Before we consider the logic of the IRT models, however, we need to introduce a suitable coding system for forced-choice responses.

Coding of forced-choice responses

The system described here is standard in the Thurstonian modeling literature (e.g. Maydeu-Olivares and Bockenholt 2005). To begin, let us assume that full rankings are

obtained. Full ranking of n items can be equivalently coded as $\tilde{n} = n(n\text{-}1)/2$ pairwise comparisons. If only two items $\{i, k\}$ are being ranked, there is only one comparison, the outcome of which can be coded as a binary variable:

$$y_{\{i,k\}} = \begin{cases} 1 & \text{if item } i \text{ is preferred over item } k \\ 0 & \text{if item } k \text{ is preferred over item } i \end{cases}. \quad (2)$$

If three items are being ranked, there are three pairwise comparisons – comparison between the first and second items, between the first and third items, and between the second and third items. Here is our earlier example ranking,

item		most / least true of me	rank order
A	I am relaxed most of the time	least	3
B	I start conversations	most	1
C	I catch on to things quickly		2

and its coding through three binary variables (binary outcomes of three pairwise comparisons),

{A, B}=0 {A, C}=0 {B, C}=1.

The reader will notice that the above binary coding contains exactly the same information as the ranking and that the original rank order can be obtained from the pairwise outcomes, and vice versa. The two coding systems are equivalent; however, the binary outcomes enable modeling of pairwise preferences using Item Response Theory, as we shall see.

Blocks of any size can be coded as pairwise comparisons. For blocks of four items (A, B, C, D), six pairwise comparisons are required:

{A, B} {A, C} {A, D} {B, C} {B, D} {C, D}.

For blocks of five, there will be 10 comparisons, etc. When full rankings are obtained, every pairwise outcome will be known. When incomplete rankings are obtained, some outcomes will be unknown in which case they can be treated as missing data.

Thurstonian IRT model

The Thurstonian IRT model was introduced by Brown and Maydeu-Olivares (2011a) to enable analysis of data arising from forced-choice tests measuring multiple traits with ranking blocks of any size. The origins of this model reside within the structural equation modeling tradition. Specifically, Maydeu-Olivares (1999) proposed a method for analyzing the mean and covariance structure of paired comparisons conforming to any observed ranking pattern, as was originally suggested by Thurstone (1931) in one of his seminal papers on choice behavior. The method relates observed choices to the differences in psychological value that respondents place on stimuli. Methods using tetrachoric correlations of dichotomous choices are used to estimate these models.

Maydeu-Olivares and Böckenholt (2005) provided a full account of Thurstonian scaling methods as applied to single ranking tasks, including cases where a factorial structure may underlie choice behavior (Thurstonian factor models). Moving from a single ranking task to multiple tasks (forced-choice blocks making up a test), and reformulating Thurstonian factor models as IRT models so that respondents' scores on underlying dimension(s) could be estimated, Brown and Maydeu-Olivares (2011a) provided the first IRT model suitable for analyzing multidimensional forced-choice data. The model may be used for estimating item parameters, and estimating correlations between the latent attributes measured by the items. Once the model parameters have been estimated, individual attribute scores and their standard errors may be established.

The Thurstonian IRT model is applicable to forced-choice formats when items are ranked within blocks, either fully or partially. Ranking blocks may be of any size, e.g. consisting of two items (item-pairs), three items (triplets), four items (quads), etc. Items in each block may indicate the same or different attributes, or any mixture of the two. The model assumes a monotonic relationship between any questionnaire item and the attribute(s) it is designed to measure; i.e. the higher is the attribute score the higher is the item endorsement level for positively keyed items; the higher is the attribute score the lower is the item endorsement level for negatively keyed items (a *dominance* response process).

The remainder of this section gives a brief overview of the theory, implementation and applications of Thurstonian IRT models.

Preference decision model. Since responses to every forced-choice block are essentially rankings, suitable models for these data are models for ranking data. One of the oldest models for ranking data was proposed by Louis Thurstone (1927; 1931). In this model, preference judgments are assumed to arise from a comparison of **unobserved** *utilities* of each item. Utility is another name for the item's *psychological value* (Thurstone 1929), described as "the affect that the object calls forth" (p. 160). For illustrative purposes, we can think of utility as the extent to which the behavior described in the item corresponds to the respondent's own behavior (Brown and Maydeu-Olivares 2013). For any given item, Thurstone further assumed that its utility is normally distributed across individuals.

Thurstone argued that any preference judgment relies on comparison of the utility values attached by the respondent to the items in question. That is, for any pair of items {i, k}, the respondent will rank item i above item k if his/her utility for item i is higher:

$$\text{prefer item} \begin{cases} i & \text{if} \quad \text{utility}_i \geq \text{utility}_k \\ k & \text{if} \quad \text{utility}_i < \text{utility}_k \end{cases}. \tag{3}$$

Taking our first example, “I start conversations” was judged to be ‘very true’ of the respondent, and “I am relaxed most of the time” to be ‘somewhat untrue’ of him/her. These (single-stimulus) ratings reflected the respondent’s utility judgments. When we asked the respondent which one of the two items was *most true* of him/her, the answer required a *comparison* of the utilities. In our example, the respondent preferred “I start conversations”, presumably, because his/her utility for this item was greater.

Having adopted a convenient coding system that presents any ranking data as several pairwise comparisons with binary outcomes (preferred – not preferred), the responses to forced-choice blocks can be easily related to the utilities of items. The binary outcomes give us the link we need to Thurstone’s law of comparative judgment, so that we can rewrite the decision rules (3), this time using differences of utilities

$$y_{\{i,k\}} = \begin{cases} 1 & \text{if} \quad \text{utility}_i - \text{utility}_k \geq 0 \\ 0 & \text{if} \quad \text{utility}_i - \text{utility}_k < 0 \end{cases}. \tag{4}$$

We can use these expressions to relate binary outcomes of pairwise comparisons within each ranking block to the underlying utilities.

Measurement model for attributes. The next step in modeling forced-choice responses is postulating a model for relations between item utilities and the psychological attributes that the items are designed to measure. The Thurstonian IRT model assumes that the utilities are related to the attributes via a linear factor analysis model.

Let us assume that each item measures one attribute[ii]. Then the utilities of items depend on the latent attributes as described by a standard linear factor analysis model:

$$\text{utility}_i = \text{mean}_i + \text{loading}_i \cdot \text{attribute}_a + \text{error}_i \tag{5}$$

In psychological assessment, the utility of items are not of interest. Instead, the focus is on measurement of attributes underlying these utilities. Therefore, the aim is to relate the observed pairwise preferences to the latent attributes directly. This is done by presenting the difference of two utilities as a function of the latent attributes,

$$\text{utility}_i - \text{utility}_k =$$

$$= (\text{mean}_i - \text{mean}_k) + (\text{loading}_i \cdot \text{attribute}_a - \text{loading}_k \cdot \text{attribute}_b) + (\text{error}_i - \text{error}_k). \quad (6)$$

Looking again at the decision rule (4), which describes preference for one item or the other depending on the item utilities, it becomes clear that we can relate the binary outcome to the psychological attributes using (6). If we further simplify this relationship by replacing the difference of means with a single threshold value

$$\text{threshold}_{\{i,k\}} = -(\text{mean}_i - \text{mean}_k), \quad (7)$$

the positive outcome of pairwise comparison $y_{\{i,k\}} = 1$ (item *i* is preferred) occurs when

$$-\text{threshold}_{\{i,k\}} + (\text{loading}_i \cdot \text{attribute}_a - \text{loading}_k \cdot \text{attribute}_b) + (\text{error}_i - \text{error}_k) \geq 0. \quad (8)$$

Since the attributes are continuous and the outcomes of pairwise comparisons are discrete, the application of Thurstone's model to forced-choice items results in an IRT model. In the standard factor analysis model (5), the errors are assumed to be uncorrelated. Therefore, the error part of expression (8) has variance

$$\text{var}(\text{error}_{\{i,k\}}) = \text{var}(\text{error}_i - \text{error}_k) = \text{var}(\text{error}_i) + \text{var}(\text{error}_k), \quad (9)$$

and, finally, the conditional probability of preferring item *i* to item *k* as given by the cumulative standard normal function is:

$$P\left(y_{\{i,k\}}=1\right)=\Phi\left(\frac{-\text{threshold}_{\{i,k\}}+\text{loading}_i\cdot\text{attribute}_a-\text{loading}_k\cdot\text{attribute}_b}{\sqrt{\text{var}(\text{error}_i)+\text{var}(\text{error}_k)}}\right). \quad (10)$$

This is a formulation of the Thurstonian IRT model for pairwise comparison between items measuring two different attributes.

The probability expression (10) is called the *item response function* and can be interpreted as follows. With an increase in attribute *a*, and a decrease in attribute *b*, the probability of preferring item *i* to item *k* increases (when the items have positive factor loadings). This probability is further influenced by item properties: a) factor loadings on the attributes that the items are designed to measure; b) a threshold value governing the combination of the attributes where the items' utilities are equal; and c) the error variances of the two items. Because two attributes are being measured, the item response function defines a surface, an example of which is presented in Figure 1, in which the probability of preferring one item to another is plotted against two latent attributes.

INSERT FIGURE 1 ABOUT HERE

So far, we have assumed that two items measuring different attributes are compared. If items measuring the same attribute are being compared, the conditional probability of preferring item *i* to item *k* is given by the cumulative standard normal function

$$P\left(y_{\{i,k\}}=1\right)=\Phi\left(\frac{-\text{threshold}_{\{i,k\}}+\left(\text{loading}_i-\text{loading}_k\right)\text{attribute}_a}{\sqrt{\text{var}(\text{error}_i)+\text{var}(\text{error}_k)}}\right). \quad (11)$$

One important feature of the one-dimensional model is that a comparison between two items measuring the same trait with similar factor loadings will result in a low pairwise factor loading, and the pairwise comparison will provide very little information on the latent

attribute. Therefore, if one wants to present items measuring the same attribute in a forced-choice block, items with very different factor loadings should be used, for example an item with a positive factor loading could be compared to an item with a negative loading (Maydeu-Olivares and Brown 2010).

The item response function in the one-dimensional case defines a curve, examples of which are presented in Figure 2, in which the probability of preferring one item to another is plotted against one latent attribute. Figure 2a illustrates the case in which two items with factor loadings of opposite sign yield a highly informative comparison; Figure 2b illustrates the case where two highly discriminating items yield an uninformative comparison.

INSERT FIGURE 2 ABOUT HERE

Technical detail. To enable parameter estimation, the above IRT model for pairwise preferences is embedded in a familiar structural equation modeling (SEM) framework. All pairwise comparisons and latent traits in the questionnaire are included, resulting in a single measurement model with binary outcomes (an IRT model). The following parameters are estimated: factor loadings and error variances for all items, thresholds for all pairwise comparisons, and correlations between the latent attributes.

When block size is $n = 2$ (item pairs), the model estimates one threshold, two factor loadings and one error variance per each pairwise comparison (error variances of individual items cannot be identified). In this case, the Thurstonian IRT model is simply the two-dimensional normal ogive IRT model.

When block size is $n = 3$ (triplets), $n = 4$ (quads) or greater, not all parameters are estimated freely for each pairwise comparison. The conditional probability expression (10) illustrates the item parameters estimated in this case. As many threshold values as there are

pairwise comparisons, are estimated. As many factor loadings as there are **items** are estimated. Finally, as many error variances as there are **items** are estimated. That is, all pairwise comparisons involving the same item will have the same factor loading on the attribute measured by that item. For example, pairs $\{i, k\}$ and $\{i, q\}$ will have the same factor loading on the attribute measured by item i, loading_i. Similarly, error variances of individual items, for example $\text{var}(\text{error}_i)$, are estimated. Furthermore, it follows from (6) that errors of pairwise comparisons involving the same item will have a shared part, so that local dependencies exist among them:

$$\text{cov}(\text{error}_{\{i, k\}}, \text{error}_{\{i, q\}}) = \text{var}(\text{error}_{i}) \, . \tag{12}$$

These special features when block size is $n \geq 3$ need to be specified in the measurement model; specifically, equality constraints need be placed on factor loadings and error variances relating to the same item since these values are estimated per item, not per pairwise comparison. In this case, the Thurstonian IRT model is an extension of the normal ogive model to items presented in forced-choice blocks. Practical guidance on parameter and person score estimation in the case of forced-choice blocks of size 3 is given in the data analysis example in this chapter (see Data analysis example with the Forced-Choice Five Factor Markers).

Thurstonian IRT models can be estimated using any method but in typical applications, there are too many latent traits for maximum likelihood estimation to be feasible. The recommended alternative is to resort to limited information methods based on tetrachoric correlations. After the item parameters have been estimated, individual scores on latent attributes are estimated by the Maximum A Posteriori (MAP) method. The reader is referred to Brown and Maydeu-Olivares (2011a) for further technical detail.

These models can be estimated in Mplus (Muthén and Muthén 1998-2012), which also performs estimation of attribute scores for individuals. The data analysis example in this chapter illustrates the workings of the Thurstonian modelling approach using Mplus.

Applications. Applications of the Thurstonian IRT model so far have included re-analysis of existing forced-choice data and development of new forced-choice questionnaires. Re-analysis of existing forced-choice data is particularly interesting since it enables direct comparison between classical and model-based IRT scoring. Research using the Customer Contact Styles Questionnaire (CCSQ; SHL 1997) demonstrates that even questionnaires with challenging features such as large block size and the use of incomplete rankings can be analyzed successfully. Furthermore, individual attribute scores estimated by the IRT method are free from the problems of ipsative data (Brown and Maydeu-Olivares 2013).

Another example is the development of a short version of the Occupational Personality Questionnaire (OPQ32r; Brown 2009), in which the questionnaire's redesign was informed by the use of item response modeling, and new IRT-based scoring was applied to produce individual attribute scores. This example illustrates how IRT modeling may be applied to re-analyze data from an existing assessment tool, and how to use this information to re-develop the tool, enhancing its strong features and transforming its scoring protocol.

For another example, a short forced-choice measure of the Five Factors of personality has been developed (Forced-Choice Five Factor Markers; Brown and Maydeu-Olivares 2011b) using 60 items from the International Personality Item Pool (IPIP; Goldberg 1992). The development of this measure was informed directly by simulation studies using IRT modeling. The data analysis example in this chapter uses data collected with this questionnaire.

An interesting application outside of tests that use explicit forced-choice formats is reported in a recent study by Lang and colleagues (2012), who addressed the complexities in

measuring implicit motives using the Operant Motive Test (OMT; Kuhl and Scheffer 2002). The OMT asks respondents to generate stories in response to ambiguous pictures, assessing motives of power, affiliation and achievement. Lang and colleagues argued that the notion of utility maximization, so central to Thurstone's account of choice behaviors, applies to implicit motives too. Only the strongest implicit motive is expressed in the narrated story, and therefore the observed responses must be modeled in relation to choices between different implicit (latent) motives. This research suggests a wider applicability of Thurstonian IRT modeling to psychological data than the authors of this model first envisaged, and we look forward to new developments in this area.

Zinnes-Griggs model for unidimensional pairwise preferences

Zinnes and Griggs (1974) introduced an IRT model describing preference judgments when choosing one of two items measuring the same attribute. This model originates from Coombs's (1950) 'unfolding' scaling tradition for modeling preference data.

Preference decision model. Coombs (1950) explained people's preferences for stimuli by their relative closeness to the position on the attribute that the stimuli represent. According to this model, when facing a choice between two stimuli, the person will prefer the stimulus closer to their own position on the attribute ('ideal point'). The ideal point is the person's location on the attribute of interest – the person's attribute score, and hence

$$\text{prefer item } \begin{cases} i & \text{if} \quad \left|\text{attribute} - \text{location}_i\right| \le \left|\text{attribute} - \text{location}_k\right| \\ k & \text{if} \quad \left|\text{attribute} - \text{location}_i\right| > \left|\text{attribute} - \text{location}_k\right| \end{cases}. \tag{13}$$

For example, when choosing between statements measuring Extraversion, a respondent will choose a statement that represents a standing on the Extraversion attribute that is closest to their own. Any preferential rank ordering, therefore, can be thought of as an

ordering of the stimuli's locations 'folded' at the person's ideal point (conversely, the stimuli locations can be uncovered by 'unfolding' the rank orders – hence the name of this decision model).

Zinnes and Griggs (1974) recognized the limitations of Coombs's deterministic model and modified it by talking about noisy 'perceptions' of the locations of both the stimuli and the person's own ideal point at the time of comparison. Thus, Zinnes and Griggs considered three normally distributed random variables with expected values corresponding to the person's own ideal point, and the two items' locations.

Measurement model for attributes. The preference decision model adopted here implicitly assumes an *ideal-point* response process for every item involved in the comparisons. The term 'ideal point' was coined by Coombs (1950) based on the original work of Thurstone (1928), who described a process of responding to attitude items. Thurstone argued that the utility value for a statement such as 'Fire arms should not belong in private hands' is the highest for individuals with this exact level of attitude towards 'militarism', and reduces for persons with **both** higher and lower levels of this attitude. Coombs called this maximum preference point on the attitude continuum the individual's 'ideal point', which is characteristic of each person.

Originally suggested for attitude items, it has recently been proposed that the ideal-point models can be applied more generally (Drasgow, Chernyshenko, and Stark 2010). Items in, say, the personality domain do not have to represent an extremely positive or negative standing on the attributes of interest. Items which represent intermediate or average positions can also be presented: For instance, 'My attention to detail is about average'. For such items, the linear factor analytic model is unlikely to represent the relationship between the item utility and the underlying construct (Conscientiousness) correctly, because the utility of this

item is likely to peak around the average Conscientiousness score, and be lower for respondents with either high or low scores.

The latent tendency to endorse an item (utility in Thurstonian terms) in the Zinnes-Griggs probabilistic model is an inverse of the absolute difference between the person's attribute score and the item location, and their random errors,

$$\text{utility}_i = -\left|\left(\text{attribute} + \text{error}_a\right) - \left(\text{location}_i + \text{error}_i\right)\right|. \tag{14}$$

Because the absolute difference between the person's attribute and the item location is opposite to the person's utility for the item, Zinnes and Griggs called it '*disutility*'. They showed that the probability of preferring one item to another conditional on the person's ideal point and the items' locations (all placed on the same attribute continuum) is a one-dimensional IRT model given by a linear combination of cumulative standard normal distribution functions:

$$P\left(y_{\{i,k\}} = 1\right) = 1 - \Phi\left(a_{\{i,k\}}\right) - \Phi\left(b_{\{i,k\}}\right) + 2\Phi\left(a_{\{i,k\}}\right)\Phi\left(b_{\{i,k\}}\right), \tag{15}$$

where

$$a_{\{i,k\}} = \left(2 \cdot \text{attribute} - \text{location}_i - \text{location}_k\right) / \sqrt{3}$$
$$b_{\{i,k\}} = \text{location}_i - \text{location}_k$$

It can be seen that the conditional probability depends only on the person's score and the item locations. The model assumes that the items vary only in their locations on the underlying attribute continuum, thus all items are assumed to be **equally discriminating**. The item response function for a pairwise comparison can be plotted against the latent attribute, as illustrated in Figure 3. This figure shows that when items with similar locations are compared, the comparison yields a very 'flat' function with a shallow slope; and when items with very dissimilar locations are compared, the comparison yields a function with a steep

slope. Therefore, comparisons between items located closely on the same attribute are non-informative. This is a very similar feature to the one we observed when comparing items with similar factor loadings under the Thurstonian IRT model.

INSERT FIGURE 3 ABOUT HERE

Technical detail. Item parameters (locations) can be estimated by the marginal maximum likelihood (MML) procedure (Stark and Drasgow 2002). Person parameters (attribute scores) can be estimated by either Expected a Posteriori (EAP) or Maximum a Posteriori (MAP) methods; EAP is recommended because it is a non-iterative procedure and is fast and efficient for one-dimensional models.

Applications. This straightforward unidimensional model has been applied recently in workplace assessments, with an added advantage of employing computerized adaptive algorithms. Borman and colleagues (2001) used it to measure job performance via a computerized adaptive test, and Schneider and colleagues (2007) used the model to build the Navy Computerized Adaptive Personality Scales (NCAPS).

Multi-Unidimensional Pairwise-Preference (MUPP) model

The Multi-Unidimensional Pairwise-Preference (MUPP) model was first introduced by Stark (2002) to enable scoring of forced-choice tests measuring multiple traits using item-pairs (i.e. blocks of size 2). The model, further developed by Stark, Chernyshenko and Drasgow (2005), adopts yet another approach to explaining preference judgments originally suggested by Andrich (1989). Andrich proposed that instead of devising an explicit multidimensional model for pairwise comparisons, the probability of preferring one item to

another might be approximated by the joint probability of accepting one item and rejecting the other. These probabilities of acceptance and rejection are based on characteristics of each of the items involved (the items' IRT parameters) established through single-stimulus item trialing.

The MUPP model may be used to create new forced-choice tests by assembling pairs of items based on their single-stimulus IRT parameters, and to estimate individuals' scores on the latent attributes. However, the model does not allow estimating item parameters from the actual forced-choice data. Items in each pair may indicate the same or different attributes (hence the name multi-unidimensional). The MUPP model assumes an ideal point response process for the items involved in comparisons, and therefore it represents Coombs's unfolding tradition. The remainder of this section gives a brief overview of the theory, implementation and applications of MUPP models.

Preference decision model. Andrich (1989) constructed models for pairwise preferences from models for single-stimulus responses. His rationale was to consider the outcome $y_{\{i,\,k\}}=1$ as a broad endorsement of item *i* and a non-endorsement of item *k*, P(1, 0). Conversely, the outcome $y_{\{i,\,k\}}=0$ corresponds to a non-endorsement of item *i* and an endorsement of item *k*, P(0, 1). The other two possible evaluation outcomes, of either endorsing two items (1, 1) or not endorsing ether (0, 0) are not admissible in a forced-choice task. Then the conditional probability of preferring item *i* to item *k* is given by the joint probability of accepting one item and rejecting the other, divided by the total probability of the two admissible outcomes,

$$P\left(y_{\{i,k\}}=1\right)=\frac{P_{ik}\left(1,\,0\right)}{P_{ik}\left(1,\,0\right)+P_{ik}\left(0,\,1\right)}. \qquad (16)$$

Acceptances and rejections of individual items are assumed to be independent events, conditional only on the attributes the item measures, therefore

$$P_{\{i,k\}}(1) = \frac{P_i(1 \mid \text{attribute}_a) P_k(0 \mid \text{attribute}_b)}{P_i(1 \mid \text{attribute}_a) P_k(0 \mid \text{attribute}_b) + P_i(0 \mid \text{attribute}_a) P_k(1 \mid \text{attribute}_b)}. \qquad (17)$$

The measurement model for attributes. Once the probability of preferring one item to another has been established from the probabilities of endorsing and not endorsing the individual items, the latter can be easily described using any suitable unidimensional IRT models. Stark, Chernyshenko and Drasgow use ideal point models to link individual items and the attributes they measure. Specifically, they advocate the use of a binary version of the Generalized Graded Unfolding Model or GGUM (Roberts, Donoghue, and Laughlin 2000). Unlike the Zinnes-Griggs model, which assumes that all items are equally discriminating, the use of GGUM allows a much wider class of items to be used in pairwise comparisons – items measuring different attributes, having different discriminating power, different locations and even different maximum probability of endorsement.

The two-dimensional item response function (17) in conjunction with the binary GGUM defines a surface, an example of which is presented in Figure 4. The reader can see that this is a more complex surface that that given in Figure 1 for the Thurstonian IRT model, because the ideal point process of responding to individual items causes a ‘number of peaks and valleys’ (Drasgow, Chernyshenko, and Stark 2009; page 74) in the response surface. If items measuring the same dimension are used, the MUPP gives a curve similar to those depicted in Figure 3.

INSERT FIGURE 4 ABOUT HERE

Technical detail. Person parameters (attribute scores) can be estimated by the Bayes modal method (Maximum a Posteriori, or MAP). Item parameters used in the unidimensional probability expressions (17) may be estimated using the freely available GGUM program (http://www.psychology.gatech.edu/unfolding/) from data gathered in single-stimulus item trials. The item parameters and the correlations between latent attributes can only be estimated from single-stimulus data, not forced-choice data.

Applications. To date, the MUPP model has been used to create new forced-choice questionnaires with items presented in pairs, and to estimate person attribute scores, after item parameters have been estimated from single stimulus trials (e.g. Chernyshenko et al. 2009). Most recently, the MUPP model was used in the development of the Tailored Adaptive Personality Assessment System or TAPAS (Drasgow, Chernyshenko, and Stark 2010), a comprehensive application taking advantage of computerized adaptive technology to select item pairs maximizing multidimensional information for person score estimation. The TAPAS is easily customizable to measure any of the 23 personality facets deemed important for predicting job performance in civil or military organizations.

McCloy-Heggestad-Reeve unfolding model for multidimensional ranking blocks

McCloy, Heggestad and Reeve (2005) sketched a theoretical model for the process of responding to multidimensional forced-choice blocks compiled from ideal-point items, and used this model to create a system for item selection and scoring that would enable accurate estimation of latent attributes.

The preference decision model. This approach also belongs to the family of Coombs's unfolding models. The model explains preferences for one item over another by the relative distances between the item locations and the respondent's attribute scores. This is an extension of Coombs's original one-dimensional model to the multidimensional case

(Coombs 1964; also Bennett and Hayes 1960). According to this model, a person will prefer an item to the extent that the person and the item are located nearer each other in multidimensional space than is true for another person and item located elsewhere in that space.

Measurement model for attributes. Just as in the Zinnes-Griggs model, equally discriminating items with ideal-point response functions are used. However, items measuring different attributes may be compared, and forced-choice blocks can be comprised of more than 2 items. The use of equally discriminating items with ideal-point response functions is necessary for this model to yield forced-choice designs that are effective for accurate estimation of attribute scores. Once item parameters have been established, a questionnaire can be assembled from blocks of items with locations that vary across the attribute space.

Data analysis example with the Forced-Choice Five Factor Markers

This data analysis example is provided to illustrate how the Thurstonian IRT modelling approach can be used in practice for item parameter estimation and scoring individuals. Data for this example are available for download together with a questionnaire form and Mplus input files necessary to run the analyses from the handbook website. In addition to the ready-made Mplus input files for this example, the reader can also download an Excel macro, which allows building Mplus syntax for a wide range of forced-choice designs using simple steps. The macro comes with a User Guide providing step-by-step instructions.

Measure

We consider real participant data collected using the Forced-Choice Five Factor Markers questionnaire (Brown and Maydeu-Olivares 2011b). Items for the questionnaire

were drawn from the International Personality Item Pool (IPIP), more specifically from its subset of items measuring the Big Five factor markers (Goldberg 1992). The forced-choice questionnaire consists of 60 items, which are presented in 20 triplets (blocks of 3), with all items within triplets measuring different attributes (so called *multidimensional forced choice*). Participants are asked to select one "most like me" item, and one "least like me" item from each block. The first triplet from this questionnaire was presented earlier in this chapter, together with example choices:

	most / least true of me
I am relaxed most of the time	least
I start conversations	most
I catch on to things quickly	

Each trait is measured with 12 items (8 positively keyed and 4 negatively keyed). The questionnaire "key" is given in Table 1.

Sample

Four-hundred-and-thirty-eight volunteers from the UK completed the questionnaire online in return for a feedback report. The sample was balanced in terms of gender (48.4% male); age ranged from 16 to 59 years, mean = 33.3, standard deviation = 10.37 years. This is the same sample described in Brown and Maydeu-Olivares (2011a).

Coding and describing data

The forced-choice design here is full ranking using 20 blocks of $n = 3$ items. Responses to each block are coded using $\tilde{n} = n(n\text{-}1)/2 = 3$ pairwise comparisons, making $20 \times 3 = 60$ pairwise comparisons in total. The outcome of each comparison is coded either 1 or 0 according to (2), therefore the data are binary.

The data file "FCFFMdata.dat" consists of 438 rows of data, one row per participant. Each row contains an identification number (ID), and 60 binary outcomes of pairwise comparisons. Here is an extract of Mplus syntax declaring our data file and variables:

```
DATA:    FILE IS 'FCFFMdata.dat';
VARIABLE:
  Names ARE ID
                i1i2 i1i3 i2i3  !first block
                i4i5 i4i6 i5i6  !second block
                …  {the rest of pairwise comparisons go here}
                i58i59 i58i60 i59i60;
  USEVARIABLES ARE i1i2-i59i60;  !ID is not used in analysis
  AUXILIARY IS ID;    !Writes ID into file with estimated scores
  CATEGORICAL ARE ALL;  !declaring all used variables categorical
```

Setting analysis options

The Unweighted Least Squares estimator with robust standard errors (denoted ULSMV in Mplus) is used to estimate the model[iii]. The parameterisation with unstandardized thresholds and factor loadings used in the conditional probability expression (10) is denoted 'theta' in Mplus. Declaring these settings completes the ANALYSIS section.

```
ANALYSIS:
    ESTIMATOR = ulsmv;
    PARAMETERIZATION = theta;
```

It is important that the correct IRT parameterization ('theta') be specified. Once the parameters have been estimated, they can be transformed using different parameterizations. For example, the results can be standardized with respect to the error variances of pairwise comparisons (their **error** variances are set to 1), to obtain the so-called *intercept/slope* IRT

parameterization. This is done by dividing the threshold and the factor loadings of each pairwise comparison {i, *k*}, by the square root of its error variance, $\sqrt{var\left(error_{\{i,\,k\}}\right)}$.

Alternatively, the results can be standardized with respect to the variances of pairwise comparisons (their **total** variances are set to 1). This is done by requesting the standardization with respect to observed variables, typing "OUTPUT: STDY;" in Mplus.

Model setup

The first part of MODEL command in Mplus describes the hypothesized factor structure of our questionnaire. Every latent attribute is defined "BY" its indicators, the pairwise comparisons. To provide a metric for the latent attributes, their variances are set to unity. For every pairwise comparison declared under an attribute, its factor loading will be estimated. We can provide starting values for this estimation depending on the item's position in the comparison and the keyed direction (whether it was designed to be a positive indicator of an attribute or a negative indicator). Remember that the first item in comparison retains the original sign of its factor loading, while the second has this sign reversed (refer to expression (10)). To aid estimation, we can set the starting values to either 1 or -1 depending on the item's position and keying. Syntax will look as follows:

```
N  BY              !Neuroticism
       i1i2*-1  (L1)
       i1i3*-1  (L1)
       ...  {the rest of pairwise comparisons involving items measuring N}
;
E  BY              !Extraversion
     i1i2*-1  (L2_n)
     i2i3*1  (L2)
       ...  {the rest of pairwise comparisons involving items measuring E}
```

```
;
O  BY              !Openness
     i1i3*-1  (L3_n)
     i2i3*-1  (L3_n)
       ...  {the rest of pairwise comparisons involving items measuring O}
;
A  BY              !Agreeableness
     i4i5*1  (L4)
     i4i6*1  (L4)
       ...  {the rest of pairwise comparisons involving items measuring A}
;
C  BY              !Conscientiousness
     i4i5*-1  (L5_n)
     i5i6*1  (L5)
       ...  {the rest of pairwise comparisons involving items measuring C}
;
N-C@1;         ! variances for all factors are set to 1
```

Symbols in brackets inside the 'BY' commands refer to the first special feature of forced-choice triplets. This special feature is that any two pairwise comparisons in each block involve the same item. For example, comparisons {1, 2} and {1, 3} both involve item 1. As explained in Thurstonian IRT model / Technical detail section, the factor loading of item 1, "loading_1", has to be the same in both comparisons. The way to tell Mplus to constrain both loadings to be the same is to give them the same parameter name (here, L1). When the item order in two comparisons is different, for instance, item 2 is last in comparison {1, 2} and first in comparison {2, 3}, the factor loadings are the same in magnitude but opposite in sign. Hence, we give the parameters different names (L2_n and L2), but constrain them the reverse of each other:

```
MODEL CONSTRAINT:
!factor loadings relating to the same item are equal in absolute value
```

```
L2_n = -L2;
L5_n = -L5;
... {remaining constraints on factor loadings go here}
```

Similar considerations apply to error variance of pairwise comparisons involving the same item. First, according to (9), the error variance of any pairwise comparison {*i*, *k*} is the sum of two components – the error variance of item *i* and the error variance of item *k*. Second, according to (12), errors of pairwise comparisons involving the same item are not independent; instead, their covariance equals var($error_i$). A fragment of Mplus syntax specifying these relationships for the first block is provided below:

```
! declare parameters for error variances of pairwise comparisons
! and set their starting values
      i1i2*2 (e1e2);
      i1i3*2 (e1e3);
      i2i3*2 (e2e3);
      … {remaining blocks go here}
 ! specify covariances between pairwise comparisons involving the same item
 ! set parameters for error variances of items, and set their starting values
      i1i2 WITH i1i3*1 (e1);
      i1i2 WITH i2i3*-1 (e2_n);
      i1i3 WITH i2i3*1 (e3);
      … {remaining blocks go here}

 MODEL CONSTRAINT:
 ! error variance of every pairwise comparison equals the sum of item error variances
      e1e2 = e1 - e2_n;
      e1e3 = e1 + e3;
      e2e3 = -e2_n + e3;
      … {remaining blocks go here}
```

It can be seen that only three unique parameters are estimated here: e1, e2_n and e3. These are the error variances of the three items in the first triplet. The error variances of the pairwise comparisons are composites of these three parameters.

Finally, fixing the error variance of one item per block for identification (here, we arbitrarily fix the variance of the last item) completes the MODEL CONSTRAINT section:

```
MODEL CONSTRAINT:
! fix one error variance in each block for identification
        e3=1;   !first block
        e6=1;   !second block
        …  {errors for remaining blocks go here}
        e60=1;
```

The above settings and syntax may seem complex and writing them out error prone. The good news is that the researcher does not have to do any syntax writing – it is automatically written by the Excel macro based on very basic information the researcher provides. For full detail, see the User Guide supplied with the macro.

Estimating attribute scores for individuals

Attribute scores for individuals in the sample can be estimated after the model parameters have been established. This is conveniently implemented in M*plus* as an option within the estimation process, using the empirical Bayesian (MAP) estimator (Muthén, 1998-2004). The estimated scores can be saved in a separate file for further use:

```
SAVE: FILE IS 'FCFFMresults.dat';
        SAVE=FSCORES;
```

When estimating person attribute scores, the estimator makes a simplifying assumption that local independence holds. The use of this simplification for scoring individuals has little impact on the accuracy of the estimates (Maydeu-Olivares & Brown,

2010). Estimation of individual scores is very fast (scores for 438 individuals in this sample are estimated and saved in a few seconds).

Interpreting Mplus output

Estimation of our model takes little time (around a minute depending on the computer). Here we briefly discuss the most important features of Mplus output that are specific to Thurstonian IRT models.

The model estimation part of the output begins with two warnings. The first warning is "THE RESIDUAL COVARIANCE MATRIX (THETA) IS NOT POSITIVE DEFINITE. … PROBLEM INVOLVING VARIABLE I2I3." This is normal and refers to the fact that by design, the residual covariance matrix in Thurstonian models is not of full rank (Maydeu-Olivares and Böckenholt 2005). The second warning is "THE MODEL CONTAINS A NON-ZERO CORRELATION BETWEEN DEPENDENT VARIABLES. SUCH CORRELATIONS ARE IGNORED IN THE COMPUTATION OF THE FACTOR SCORES." This warns the researcher that the local dependencies that we have specified for pairwise comparisons are ignored when estimating the attribute scores for individuals.

Next, goodness of fit statistics and indices are printed. The model yields a chi-square of 2106.06 on 1660 degrees of freedom; however, the degrees of freedom have to be adjusted since there is one redundancy in every block of three items (see Brown and Maydeu-Olivares, 2012). The fit indices including the degrees of freedom in their computation also need to be adjusted. Overall, there are 20 redundancies in the model, so that the correct $df = 1640$, and the correct RMSEA = 0.025.

The rest of the output provides parameter estimates for our model. First, the factor loadings are printed. The reader can compare the below fragment of the output with the input

instructions we provided for the factor structure, and confirm that the factor loadings we constrained to be equal are indeed equal in pairwise comparisons relating to the same item (may have opposite sign depending on the item's order in the pairwise comparison):

MODEL RESULTS

			Two-Tailed	
	Estimate	S.E.	Est./S.E.	P-Value
N BY				
I1I2	-0.705	0.173	-4.086	0.000
I1I3	-0.705	0.173	-4.086	0.000
<…>				
E BY				
I1I2	-1.108	0.192	-5.762	0.000
I2I3	1.108	0.192	5.762	0.000
<…>				
O BY				
I1I3	-1.024	0.202	-5.060	0.000
I2I3	-1.024	0.202	-5.060	0.000
<…>				
A BY				
I4I5	0.844	0.139	6.070	0.000
I4I6	0.844	0.139	6.070	0.000
<…>				
C BY				
I4I5	-0.994	0.139	-7.124	0.000
I5I6	0.994	0.139	7.124	0.000
<…>				

Some of the factor loadings are above 1. This is because these are **unstandardized** parameters, as defined by (10).

Next, the factor covariances are printed. Since we set the factor variances to 1, these are correlations between the latent attributes:

			Two-Tailed	
	Estimate	S.E.	Est./S.E.	P-Value
E WITH				
N	-0.404	0.060	-6.693	0.000
O WITH				
N	-0.482	0.068	-7.068	0.000
E	0.479	0.061	7.867	0.000

A	WITH				
	N	-0.403	0.075	-5.372	0.000
	E	0.413	0.068	6.100	0.000
	O	0.145	0.086	1.681	0.093
C	WITH				
	N	-0.299	0.073	-4.088	0.000
	E	0.232	0.072	3.205	0.001
	O	0.349	0.070	5.019	0.000
	A	0.307	0.076	4.044	0.000

Next, the estimated covariances between the errors of the pairwise comparisons are printed. The reader is reminded that these covariances equal to the error variance of the common item involved in both comparisons, as we specified, and hence the covariance between the error of comparison {1, 2} (referred to as I1I2 in Mplus) and {1, 3} (referred to as I1I3) is simply the error variance of item 1, var($error_1$). A fragment below relates to the first block:

				Two-Tailed	
		Estimate	S.E.	Est./S.E.	P-Value
I1I2	WITH				
	I1I3	1.463	0.517	2.831	0.005
	I2I3	-0.242	0.221	-1.098	0.272
I1I3	WITH				
	I2I3	1.000	0.000	Infinity	0.000

Because we fixed the error for the last item, item 3, to unity, no standard error was estimated for this parameter. The covariance between errors of {1, 2} and {2, 3} is negative since item 2 is first in comparison {2, 3} and last in comparison {1, 2}; however, the error variance of item 2 is of course positive, 0.242. The reader can compare the above output with the error variances of the pairwise comparisons printed further in the output:

			Two-Tailed	
	Estimate	S.E.	Est./S.E.	P-Value
Residual Variances				
I1I2	1.705	0.619	2.754	0.006
I1I3	2.463	0.517	4.766	0.000
I2I3	1.242	0.221	5.628	0.000

It can be seen that the error variances of the pairwise comparisons are the sums of error variances of the two items involved in comparison, for instance 1.705 = 1.463 + 0.242. This is exactly how we specified the relationships between errors. Thus, it is clear that the only unique estimable parameters are the item errors – the errors of pairwise comparisons are simply their combinations.

Finally, the thresholds for each pairwise comparison are printed. The symbol $1 refers to the first and the only threshold, since the data are binary. Here are the thresholds for the first block.

			Two-Tailed	
	Estimate	S.E.	Est./S.E.	P-Value
Thresholds				
I1I2$1	0.231	0.108	2.136	0.033
I1I3$1	1.287	0.192	6.706	0.000
I2I3$1	1.161	0.161	7.227	0.000

The estimated thresholds, factor loadings and error variances for these data are given in Table 1. The estimated correlations between the five attributes are given in Table 2. They are provided so that the reader can practice finding the required parameters in Mplus outputs. For further practical guidance on model setup and identification for different forced-choice designs, including treatment of missing data in incomplete rankings, and example Mplus syntax, see Brown and Maydeu-Olivares (2012).

Recommendations for creating effective forced-choice assessments

Recent methodological and technological advances have made item response modeling of forced-choice data possible, and new approaches to creating and scoring forced-choice tests have emerged. However, given that response processes involve comparisons between two or more items, and generally lead to multidimensional IRT models, test development using forced-choice formats is a more complex endeavor than when single-stimulus formats are used.

A number of factors affect forced-choice test design decisions. We have already discussed how the nature of preferential choice dictates rules for selecting items in one-dimensional comparisons. To yield informative comparisons, items with very different factor loadings must be used under factor analysis models, and items with very different locations should be used under ideal-point models. These are not limitations of particular response models – these are the limitations to comparative judgments in recovering absolute information. For instance, a small proportion (5-10%) of unidimensional item-pairs have been recommended for use alongside multidimensional item pairs to identify the latent trait metric under the MUPP model (Drasgow, Chernyshenko, and Stark 2009). Under Thurstonian IRT modeling, the latent attribute metric is generally identified without any unidimensional comparisons (Brown and Maydeu-Olivares 2012), and therefore this model enables the use of purely multidimensional forced-choice formats.

Brown and Maydeu-Olivares (2011a) provide guidelines for constructing ordinal forced-choice questionnaires with common dominance items that are effective in measuring multiple attributes. They show that in questionnaires containing multidimensional comparisons only, the attribute scores can be estimated accurately if sufficient numbers of good quality items are used, and the following rules are met:

Keyed direction of items. When forced-choice tests involve comparisons between items keyed in the same direction, as well as items keyed in the opposite direction, accurate estimation of attribute scores can be achieved with any number of traits, and any level of trait correlations. However, when all comparisons are between items keyed in the same direction, the quality of measurement depends on the number of attributes assessed in the test.

Number of attributes. It is possible to estimate attribute scores accurately using only positively keyed items, when the number of traits assessed is large (20-30 or more) and the attributes are largely independent on average.

Correlations between traits. Given the same number of attributes, the lower the average correlation between them the more accurate the score estimation will be.

Block size. Given the same number of items available, combining them in larger blocks increases the level of information each item provides for latent attribute estimation. This is because the number of pairwise comparisons increases rapidly as the number of items in a block increases. However, increasing block size increases cognitive complexity and contributes to respondents' fatigue and random responding.

The reader interested in using a forced-choice format must remember that the format itself cannot correct for faults in item writing, and in some ways makes these faults more apparent. In our experience, using negations in forced choice formats is more problematic than in single-stimulus formats. Item length can also be a problem, particularly when four or more items are compared in one block. Adhering to good item writing practice (Brown and Maydeu-Olivares 2010) and the above general rules when designing a forced-choice questionnaire will increase its measurement quality.

To the authors' knowledge, all item response modeling of forced-choice questionnaires to date has been conducted with pairwise comparisons and ranking blocks.

Pairwise comparisons using graded preferences, and compositional forced-choice formats are yet to be explored, therefore more research on forced-choice questionnaire design is needed.

Directions for future research and concluding remarks

Thanks to the recent developments in IRT-based analysis and scoring, forced-choice tests are enjoying a lot of attention from psychometricians, test developers and test users alike. They are gaining popularity in workplace assessments, where concerns about response distortions are strong. We, however, would like to draw the reader's attention to applications that have been overlooked in the forced-choice literature, and where we believe the gains might be very important to the advancement of science.

In **cross-cultural research**, where culture-specific response sets present a challenge for comparability of scores (Van Herk et al. 2004; Johnson, Kulesa, Cho, and Shavitt 2005), the use of forced-choice formats are bound to be advantageous. A recent example of cross-cultural personality research using forced-choice tests is represented by a study of Bartram's (2013), which examined personality profiles across 31 cultures in relation to country-level cultural dimensions. Furthermore, measurement invariance of forced-choice tests can be formally tested; for example, constraining the thresholds, the factor loadings and the correlations between the latent attributes to be equal across two or more groups in the corresponding multi-group Thurstonian IRT model.

The use of forced-choice formats is likely to prove beneficial in **assessments of other individuals** (as in 360-degree feedback), **organizations or services** (as in satisfaction surveys). In these contexts, rater biases such as leniency/severity and the halo effects are notorious (e.g. Brown et al. 2012) and forced-choice formats are ideally placed to counter such biases and deliver more usable data.

Historically, the psychometric assessment industry and test users have been more enthusiastic about the use of forced-choice assessments than academics have been. While the former groups have been excited by the advantages in overcoming common response biases and enhanced differentiation between stimuli, the latter group has been concerned about the psychometric properties of ipsative data, particularly its interpersonal incomparability. As we have shown in this article, academic concerns about the use of ipsative data are well founded. Therefore, we advocate the use of item response modeling with forced-choice data, which, in conjunction with good test development practices has the potential to overcome these problems, and consequently, forced-choice formats have the potential to compete with single-stimulus formats in applications.

References

Andrich, David. 1989. "A probabilistic IRT model for unfolding preference data." *Applied Psychological Measurement*, 13: 193-216.

Baron, Helen. 1996. "Strength and limitations of ipsative measurement". *Journal of Occupational and Organizational Psychology*, 69: 49-56.

Bartram, Dave. 2007. "Increasing validity with forced-choice criterion measurement formats". *International Journal of Selection and Assessment*, 15: 263-272.

Bartram, Dave. 2013. "Scalar Equivalence of OPQ32: Big Five Profiles of 31 Countries". *Journal of Cross-Cultural Psychology*, 44: 61-83.

Bennett, Joseph F., and William L. Hays. 1960. "Multidimensional unfolding: Determining the dimensionality of ranked preference data." *Psychometrika*, 25, no. 1: 27-43.

Block, Jack. 1961. "*The Q-sort method in personality assessment and psychiatric research*". Springfield, IL: Charles C. Thomas.

Block, Jack. 2008. "*The Q-sort in character appraisal: Encoding subjective impressions of persons quantitatively*". Washington, DC: American Psychological Association.

Borman, Walter C., Daren E. Buck, Mary Ann Hanson, Stephan J. Motowidlo, Stephen Stark, and Fritz Drasgow. 2001. "An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales." *Journal of Applied Psychology*, 86, no. 5: 965-973.

Brown, Anna, and Dave Bartram. (2009, April). "*Doing less but getting more: Improving forced-choice measures with IRT"*. Paper presented at the 24th conference of the Society for Industrial and Organizational Psychology, New Orleans, LA. Accessed 1 July www.shl.com/assets/resources/Presentation-2009-Doing-less-but-getting-more-SIOP.pdf

Brown, Anna, Tamsin Ford, Jessica Deighton, and Miranda Wolpert. 2012. "Satisfaction in child and adolescent mental health services: Translating users' feedback into measurement." *Administration and Policy in Mental Health and Mental Health Services Research*: advance online publication.

Brown, Anna, and Alberto Maydeu-Olivares. 2010. "Issues that should not be overlooked in the dominance versus ideal point controversy." *Industrial and Organizational Psychology* 3, no. 4: 489-493.

Brown, Anna, and Alberto Maydeu-Olivares. 2011a. "Item response modeling of forced-choice questionnaires". *Educational and Psychological Measurement,* 71: 460-502. DOI: 10.1177/0013164410375112

Brown, Anna, and Alberto Maydeu-Olivares. 2011b. "Forced-choice Five Factor markers". *PsycTESTS*. DOI:10.1037/t05430-000

Brown, Anna, and Alberto Maydeu-Olivares. 2012. “Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, 44: 1135–1147. DOI: 10.3758/s13428-012-0217-x

Brown, Anna, and Alberto Maydeu-Olivares. 2013. “How IRT can solve problems of ipsative data in forced-choice questionnaires”. *Psychological Methods*, 18(1): 36-52. DOI: 10.1037/a0030641

Chan, Wai. 2003. “Analyzing ipsative data in psychological research”. *Behaviormetrika*, 30: 99-121.

Cheung, Mike W.L, and Wai Chan, W. 2002. “Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis”. *Structural Equation Modeling*, 9: 55-77. DOI: 10.1207/S15328007SEM0901_4

Christiansen, Neil D., Gary N. Burns, and George E. Montgomery. "Reconsidering forced-choice item formats for applicant personality assessment." *Human Performance*, 18, no. 3 (2005): 267-307. DOI: 10.1207/s15327043hup1803_4

Clemans, William V. 1966. “*An Analytical and Empirical Examination of Some Properties of Ipsative Measures”.* Psychometric Monograph No. 14. Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN14.pdf

Closs, S. J. 1996. “On the factoring and interpretation of ipsative data”. *Journal of Occupational Psychology*, 69: 41-47.

Coombs, Clyde H. 1950. “Psychological scaling without a unit of measurement”. *Psychological Review*, 57: 145-158.

Coombs, Clyde H. 1964. “*The theory of data”*. New York: Wiley.

Cornwell, John M., and William P. Dunlap. 1994. "On the questionable soundness of factoring ipsative data: A response to Saville & Willson." *Journal of Occupational and Organizational Psychology*, 67, no. 2: 89-100.

Cubiks. 2010. “PAPI: Personality and Preference Inventory”. Accessed 1 July http://www.cubiks.com/PRODUCTS/PERSONALITYASSESSMENTS/Pages/papi.aspx

Drasgow, Fritz, Oleksandr S. Chernyshenko, and Stephen Stark. 2009. “Test theory and personality measurement”. In *Oxford Handbook of Personality Assessment* edited by John N. Butcher, 59-80. London: Oxford University Press.

Drasgow, Fritz, Oleksandr S. Chernyshenko, and Stephen Stark. 2010. “75 years after Likert: Thurstone was right!”. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 3: 465–476.

Drasgow, Fritz, Oleksandr S. Chernyshenko, and Stephen Stark. 2010. "*Tailored Adaptive Personality Assessment System (TAPAS)*." Urbana, IL: Drasgow Consulting Group.

Dunlap, William P., and John M. Cornwell. 1994. "Factor analysis of ipsative measures." *Multivariate Behavioral Research*, 29, no. 1: 115-126.

Feldman, Marvin J., and Norman L. Corah. 1960. "Social desirability and the forced choice method." *Journal of Consulting Psychology*, 24, no. 6: 480-482.

Friedman, Hershey H., and Taiwo Amoo. 1999. "Rating the rating scales." *Journal of Marketing Management*, 9, no. 3: 114-123.

Funder, David C., R. Michael Furr, and C. Randall Colvin. 2000. "The Riverside Behavioral Q-sort: A Tool for the Description of Social Behavior." *Journal of Personality*, 68, no. 3: 451-489.

Goldberg, Lewis R. 1992. “The development of markers for the Big-Five factor structure”. *Psychological Assessment*, 4: 26-42.

Gordon, L.V. 1976. “*Survey of interpersonal values. Revised manual”*. Chicago, IL: Science Research Associates.

Gordon, L.V. 1993. "*Manual: Gordon Personal Profile-Inventory"*. San Antonio, TX: The Psychological Corporation.

Heggestad, Eric D., Morgan Morrison, Charlie L. Reeve, and Rodney A. McCloy. 2006. "Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance." *Journal of Applied Psychology*, 91, no. 1: 9-24.

Hicks, Lou E. 1970. "Some properties of ipsative, normative, and forced-choice normative measures". *Psychological Bulletin*, 74: 167-184.

Jackson, Douglas N., Victor R. Wroblewski, and Michael C. Ashton. 2000. "The impact of faking on employment tests: does forced choice offer a solution?" *Human Performance*, 13, no. 4: 371-388.

Johnson, Timothy, Patrick Kulesa, Young Ik Cho, and Sharon Shavitt. 2005. "The relation between culture and response styles evidence from 19 countries." *Journal of Cross-cultural psychology*, 36, no. 2: 264-277.

Johnson, Charles E., Robert Wood, and S. F. Blinkhorn. 1988. "Spuriouser and spuriouser: The use of ipsative personality tests." *Journal of Occupational Psychology*, 61, no. 2: 153-162.

Kolb, Alice Y., and David A. Kolb. 2005. "*The Kolb learning style inventory–version 3.1 2005 technical specifications*." Boston, MA: Hay Resource Direct.

Kuhl, Julius and David Scheffer. 2002. "*Der operante Multi-Motiv-Test (OMT): Manual [The operant multi-motive-test (OMT): Manual]*". Osnabrück, Germany: University of Osnabrück.

Kuncel, Nathan R., Lewis R. Goldberg, and Tom Kiger. 2011. "A Plea for Process in Personality Prevarication". *Human Performance*, 24 (4): 373-378.

Lang, Jonas W. B., Ingo Zettler, Christian Ewen, and Ute R. Hülsheger. 2012, August 6. "Implicit Motives, Explicit Traits, and Task and Contextual Performance at Work". *Journal of Applied Psychology*. Advance online publication. DOI: 10.1037/a0029556

Martin, Beth A., Chieh-Chen Bowen, and Steven T. Hunt. 2002. "How effective are people at faking on personality questionnaires?" *Personality and Individual Differences*, 32(2): 247-256.

Maydeu-Olivares, Alberto. 1999. "Thurstonian modeling of ranking data via mean and covariance structure analysis". *Psychometrika*, 64: 325-340.

Maydeu-Olivares, Alberto, and Ulf Böckenholt. 2005. "Structural equation modeling of paired-comparison and ranking data". *Psychological Methods*, 10: 285-304.

Maydeu-Olivares, Alberto, and Ulf Böckenholt. 2008. "Modeling subjective health outcomes: Top 10 reasons to use Thurstone's method". *Medical Care*, 46: 346-348.

Maydeu-Olivares, Alberto, and Anna Brown. 2010. "Item response modeling of paired comparison and ranking data". *Multivariate Behavioral Research*, 45: 935 - 974. DOI: 10.1080/00273171.2010.531231

McCloy, Rodney A., Eric D. Heggestad, and Charlie L. Reeve. 2005. "A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items." *Organizational Research Methods*, 8, no. 2: 222-248.

Meade, Adam W. 2004. "Psychometric problems and issues involved with creating and using ipsative measures for selection". *Journal of Occupational and Organisational Psychology*, 77: 531-552.

Mellenbergh, Gideon J. 2011. "*A conceptual introduction to psychometrics: development, analysis and application of psychological and educational tests*". Eleven International Publishing.

Muthén, Bengt O. 1998-2004. "*Mplus technical appendices"*. Los Angeles, CA: Muthén & Muthén.

Muthén, Linda K., and Bengt O. Muthén. 1998-2012. "*Mplus User's guide. Seventh edition"*. Los Angeles, CA: Muthén & Muthén.

Roberts, James S., John R. Donoghue, and James E. Laughlin. 2000. "A general item response theory model for unfolding unidimensional polytomous responses." *Applied Psychological Measurement*, 24, no. 1: 3-32.

Schneider, Robert J., Kerri L. Ferstl, Janis S. Houston, Walter C. Borman, Ronald M. Bearden, and Amanda O. Lords. 2007 "*Revision and expansion of Navy Computer Adaptive Personality Scales (NCAPS)"*. Tampa, FL: Personnel decisions research Institute Inc.

Schwarz, Norbert, Bärbel Knäuper, Hans-J. Hippler, Elisabeth Noelle-Neumann, and Leslie Clark. 1991. "Rating scales numeric values may change the meaning of scale labels." *Public Opinion Quarterly* 55, no. 4: 570-582.

SHL. 1997. "*Customer Contact: Manual and User's Guide*". Surrey, UK: SHL Group.

SHL. 2006. "*OPQ32 Technical Manual"*. Surrey, UK: SHL Group.

Stark, Stephen. 2002. "*A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment"*. Unpublished doctoral dissertation. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Stark, Stephen, Oleksandr S. Chernyshenko, and Fritz Drasgow. 2005. „An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The Multi-Unidimensional Pairwise-Preference Model". *Applied Psychological Measurement*, 29: 184-203.

Stark, Stephen, and Drasgow, Fritz. 2002. “An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model”. *Applied Psychological Measurement*, 26: 208–227.

Tenopyr, Mary L. 1988. “Artifactual reliability of forced-choice scales”. *Journal of Applied Psychology*, 73: 749-751.

Thorndike, Edward L. 1920. "A constant error in psychological ratings." *Journal of applied psychology* 4, no. 1: 25-29.

Thurstone, Louis L. 1927. “A law of comparative judgment”. *Psychological Review*, 34: 273-286.

Thurstone, Louis L. 1929. “The measurement of psychological value”. In *Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago*, edited by Thomas Vernor Smith and William Kelley Wright, 157-174. Chicago: Open Court.

Thurstone, Louis L. 1931. “Rank order as a psychophysical method”. *Journal of Experimental Psychology*, 14: 187-201.

Torgeson, Warren S. 1958. “*Theory and methods of scaling*”. New York: Wiley.

Van Herk, Hester, Ype H. Poortinga, and Theo MM Verhallen. 2004. "Response Styles in Rating Scales Evidence of Method Bias in Data From Six EU Countries." *Journal of Cross-Cultural Psychology*, 35, no. 3: 346-360.

Vasilopoulos, Nicholas L., Jeffrey M. Cucina, Natalia V. Dyomina, Courtney L. Morewitz, and Richard R. Reilly. 2006. "Forced-choice personality tests: A measure of personality and cognitive ability?." *Human Performance*, 19, no. 3: 175-199.

Wagerman, Seth A., and David C. Funder, D. C. 2009. “Personality psychology of situations”. In *Cambridge Handbook of Personality Psychology*, edited by Corr, Philip J., and Gerald Matthews, 27-42. Cambridge: Cambridge University Press.

Zerbe, Wilfred J., and Delroy L. Paulhus. 1987. "Socially desirable responding in organizational behavior: A reconception." *Academy of Management Review*, 12, no. 2: 250-264.

Zinnes, Joseph L., and Richard A. Griggs. 1974. "Probabilistic, multidimensional unfolding analysis." *Psychometrika*, 39, no. 3: 327-350.

Table 1

Estimated item parameters for the Forced-Choice Five Factor Markers example

Blocks	Items				Pairwise comparisons	
#	#	Attribute	loading	var(error)	#	threshold
1	1	N	-0.705	1.463	{1,2}	0.231
	2	E	1.108	0.242	{1,3}	1.287
	3	O	1.024	1	{2,3}	1.161
2	4	A	0.845	1.004	{4,5}	-0.327
	5	C	0.994	0.287	{4,6}	-1.237
	6	N	-0.804	1	{5,6}	-0.834
3	7	O	-0.476	0.052	{7,8}	0.203
	8	E	0.624	2.904	{7,9}	1.999
	9	A	0.822	1	{8,9}	2.102
4	10	C	0.734	1.083	{10,11}	0.415
	11	O	0.974	0.466	{10,12}	-2.155
	12	N	0.722	1	{11,12}	-2.507
5	13	A	0.552	1.249	{13,14}	-2.144
	14	N	1.194	3.157	{13,15}	-0.151
	15	E	1.243	1	{14,15}	1.813
6	16	O	0.903	0.92	{16,17}	-1.803
	17	E	-0.719	1.091	{16,18}	-2.514
	18	C	-0.720	1	{17,18}	-0.71
7	19	E	-1.251	2.711	{19,20}	-1.575
	20	N	1.864	4.828	{19,21}	2.378
	21	A	0.607	1	{20,21}	4.524
8	22	C	0.667	0.56	{22,23}	0.034
	23	O	0.665	0.287	{22,24}	-1.462
	24	E	-0.698	1	{23,24}	-1.505
9	25	O	1.235	4.803	{25,26}	-2.922
	26	N	1.379	2.276	{25,27}	-3.404
	27	A	-1.116	1	{26,27}	-0.499
10	28	C	-0.821	0.629	{28,29}	0.358
	29	N	0.636	0.675	{28,30}	1.636
	30	E	1.141	1	{29,30}	1.408
11	31	E	0.847	0.458	{31,32}	0.29
	32	A	0.676	0.42	{31,33}	0.165
	33	C	0.838	1	{32,33}	-0.249
12	34	N	-0.455	1.219	{34,35}	0.629
	35	A	0.570	0.79	{34,36}	-1.132
	36	O	-0.787	1	{35,36}	-1.887

13	37	E	-0.830	0.731	{37,38}	-0.673
	38	N	1.004	0.756	{37,39}	-0.598
	39	C	-0.985	1	{38,39}	-0.03
14	40	A	0.798	1.356	{40,41}	0.112
	41	C	1.114	1.134	{40,42}	0.301
	42	O	1.158	1	{41,42}	0.486
15	43	E	0.838	0.937	{43,44}	-0.178
	44	O	0.983	0.855	{43,45}	-2.307
	45	N	1.109	1	{44,45}	-2.424
16	46	C	1.202	2.074	{46,47}	-1.192
	47	N	-1.115	3.062	{46,48}	-3.287
	48	A	-0.417	1	{47,48}	-2.363
17	49	C	-0.830	2.558	{49,50}	2.644
	50	A	0.782	1.696	{49,51}	3.038
	51	O	0.878	1	{50,51}	0.775
18	52	A	-0.899	3.309	{52,53}	2.844
	53	E	0.956	3.941	{52,54}	-0.227
	54	O	-0.702	1	{53,54}	-2.758
19	55	O	-0.463	1.823	{55,56}	1.929
	56	C	0.801	1.036	{55,57}	-0.683
	57	N	0.701	1	{56,57}	-2.444
20	58	C	0.546	1.167	{58,59}	-1.992
	59	A	-0.339	0.315	{58,60}	-0.031
	60	E	1.052	1	{59,60}	2.027

Note: N = 438. N = Neuroticism, E = Extraversion, O = Openness, A = Agreeableness, C = Conscientiousness. Error variance of last item in each block is set to 1 for identification.

Table 2

Estimated latent attribute correlations for the Forced-Choice Five Factor Markers

	N	E	O	A	C
Neuroticism (N)	1				
Extraversion (E)	-.404**	1			
Openness (O)	-.482**	.479**	1		
Agreeableness (A)	-.403**	.413**	.145	1	
Conscientiousness (C)	-.299**	.232*	.349**	.307**	1

Note: $N = 438$. * Correlation is significant at $p < .01$; ** correlation is significant at $p < .001$.

Figures

Figure 1. Example Thurstonian item response function for preferring an item measuring attribute *a* to an item measuring attribute *b*.

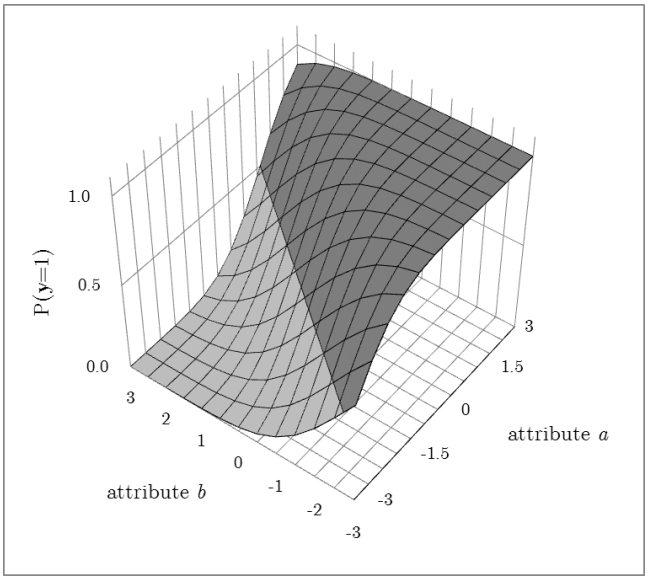
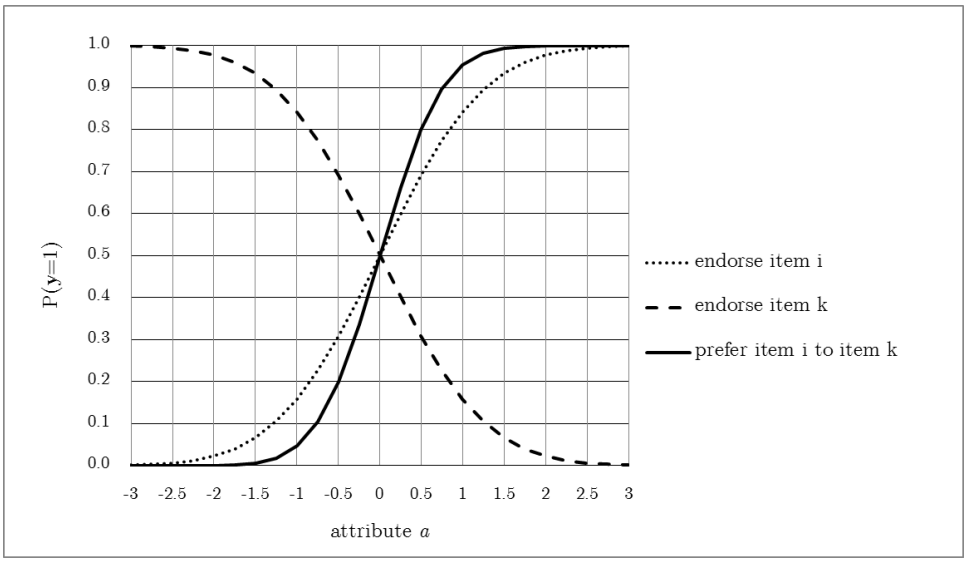
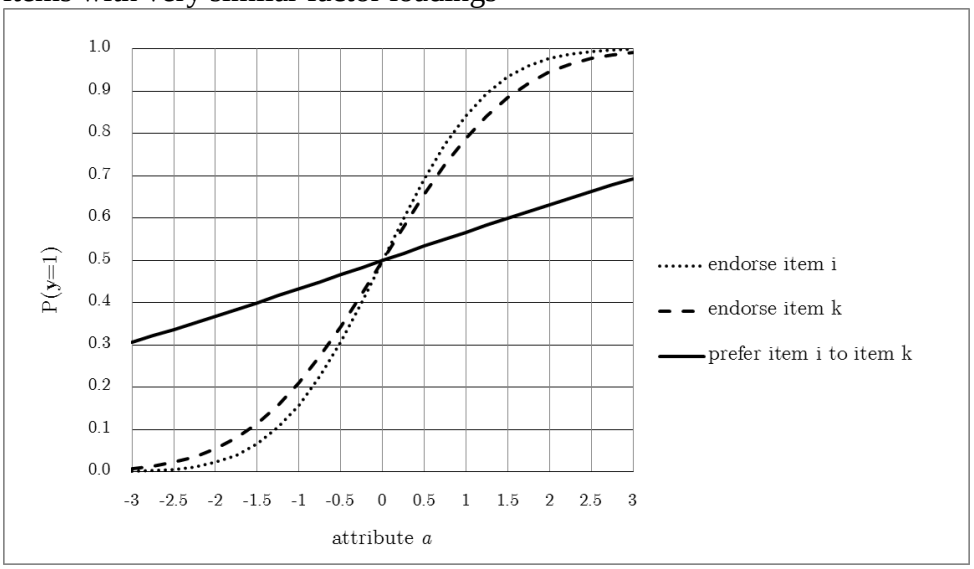


Figure 2. Example Thurstonian item response functions for preferring item *i* to item *k*; both items measure the same attribute *a*.

(a) Items with very different factor loadings

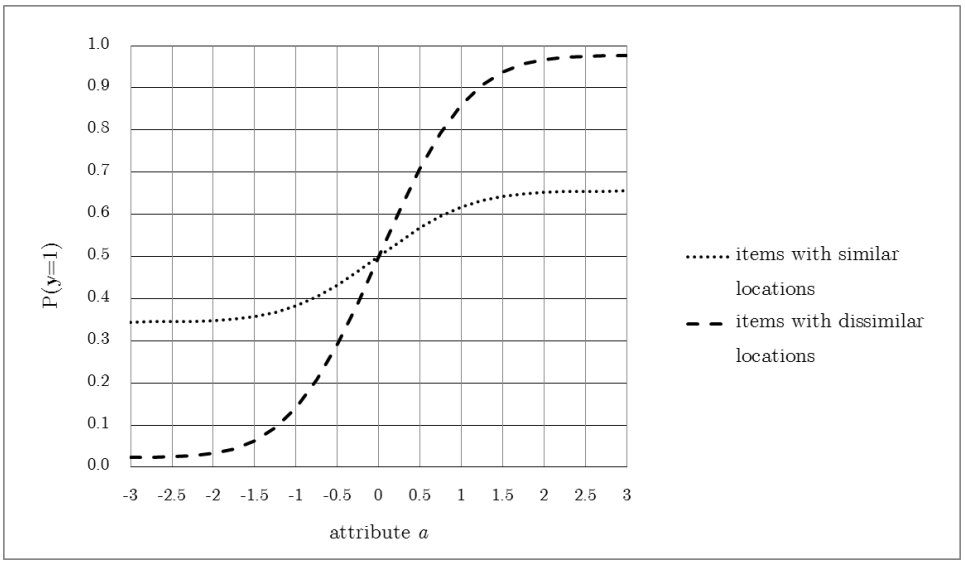


(b) items with very similar factor loadings



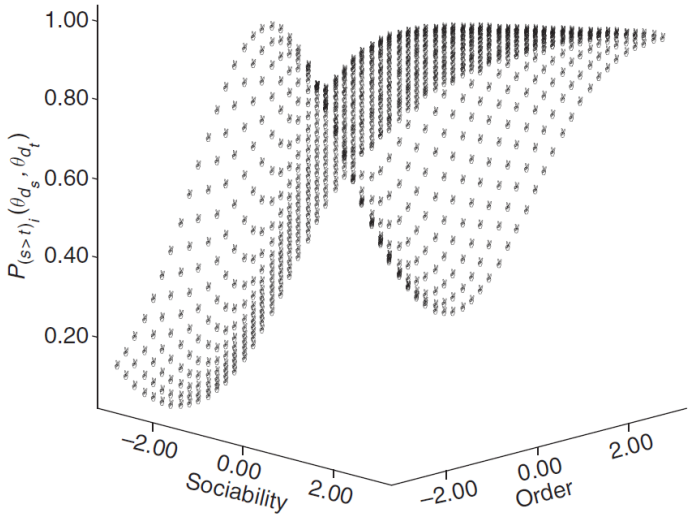
Note. The dotted and dashed lines illustrate the conditional probabilities of endorsing item *i* and *k*, respectively. The solid line illustrates the conditional probability of preferring item *i* to item *k*.

Figure 3. Example Zinnes-Griggs item response functions for preferring item *i* to item *k*; both items measure the same attribute *a*.



Note. The dotted line illustrates comparison between items with locations 0.2 and −0.2; the dashed line illustrates comparison between items with locations 1 and −1.

Figure 4. Example MUPP item response function for a pair of items measuring order and sociability.



Source: Drasgow, F., Chernyshenko, O. S., & Stark, S. (2009). Test theory and personality measurement. In J.N. Butcher (Ed.). Oxford Handbook of Personality Assessment (pp. 59-80). London: Oxford University Press.

[i] Type I (Preferential Choice) and Type III (Stimulus Comparison) can use exactly the same response formats, but differ in whether the data capture information on the respondents themselves (and whether this information is of interest). In Type I comparisons, the participant indicates own relationship with the stimuli (preference, applicability to self, etc.), and therefore can be placed on the same psychological continuum as the stimuli. On contrary, Type III data captures comparisons between stimuli regarding a certain property (brightness, sweetness, competence, fairness, etc.). In such comparisons, the respondents act merely as judges, they do not indicate own relation to the psychological continuum of interest and therefore cannot be placed on this continuum (Coombs, 1964). The Type III data is very common in marketing; it is also popular in psychometric applications where judgements about other individuals are collected.

[ii] This is certainly the aim in most forced-choice questionnaires. The Thurstonian IRT model can easily accommodate items measuring multiple attributes (see Brown and Maydeu-Olivares 2012), but for simplicity of illustration here we limit the number of measured attributes to one.

[iii] Alternatively, the Diagonally Weighted Least Squares estimator with robust standard errors may be used (denoted WLSMV in Mplus).